WHITE PAPER

BEYOND CLASSES OF SERVICE – A USER GUIDE

EXECUTIVE SUMMARY

In the past WANs were used for a small number of information and business applications, (financial accounting, supply chain, etc.); now, however, they are available to every level of the enterprise.

The flexibility of IP protocols allows the network to be shared by numerous applications. Each of these applications impacts the way a company is run and they require very differing performance levels: transactional traffic is impacted by average delay, large file transfers are bandwidth hungry, while voice traffic is very sensitive to jitter. This increasing number of different uses leads to greater network requirements along with the need to align network resources with applications requirements.

Faced with the successful deployment of private networks (IP VPNs) and the need to differentiate the performance of different applications, Service Operators are offering their clients Classes of Service (CoS) over MPLS networks.

Even though CoS is a step in the right direction, it is often not able to respond to all the issues associated with multiple application deployments on the same network. It has some natural inflexibilities and limitations.

Ipanema has gone beyond the limitations of CoS and has created the 'Business Network Optimization' solution, allowing you to align application performance requirements with the enterprise's network resources in a more specific, flexible and dynamic way.

ipanema Technologies

E-mail : info@ipanematech.com
http://www.ipanematech.com

THE NEED TO MAXIMIZE APPLICATION PERFORMANCE

Several factors explain why now, more than ever, maximizing application performance over the network by optimizing network resources is a critical task

1. **The network has become the critical infrastructure**
2. **There are a wide variety of application behaviors**
3. **The cost of poor quality is an issue**

1. The network has become the critical infrastructure

There are an ever increasing number of applications running over Wide Area Networks. Improved Web-based human interfaces require larger data transfers and more turns through the network. Performance sensitive applications such as Citrix or VoIP are on the rise. It is estimated that two thirds of application response time is now driven by the network.

1.1 Lower employee productivity

Poor application response time has a negative impact on result in irate customers abandoning baskets full of goods. At bank, a study showed that there was a direct correlation between banking applications response time and the length of the customer queues at the teller. Long queues result in unhappy customers that are open to the competition.

1.2 Lower employee productivity

Poor application response time has a negative impact on employee's morale, attitude and productivity.

A critical application response time beyond 2 seconds generates wasted user time that can translate into a 30% productivity loss. As the delay exceeds 4 seconds, users start complaining to the helpdesk, thus generating further costs. If the delay exceeds 8 seconds users may stop using the application or work overtime outside of peak hours.

2. There are a wide variety of application behaviors

The network is carrying an ever-increasing number of applications with different behaviors and needs. Transactional (SAP) and thin-client (Citrix) applications, often critical for an organization, exchange little data, but are very sensitive to transfer delays. On the other hand, real-time flows such as VoIP require low jitter. Finally, file transfers (e-mail, NetBios, ftp...), while usually less critical, are not delay sensitive but typically require large amounts of bandwidth.

TCP, the standard protocol on IP networks, does not distinguish between applications. It allows each flow to take as much available bandwidth as possible. A few bandwidth hungry flows (e-mail with attachments, print files...) can therefore upset a large number of transactional flows that require less

bandwidth, but whose performance is more critical for the organization. Critical applications from a business standpoint are not protected.

3. The cost of poor quality is an issue

Poor network quality and application performance represent a high cost for organizations. Network downtime (blackout) or, even more so, non-performance (brownout) leads to losses in revenue and productivity. For instance, according to a 2004 survey from the Standish Group, the cost per minute of downtime for an ERP system for a Fortune 1000 company is $13 000.

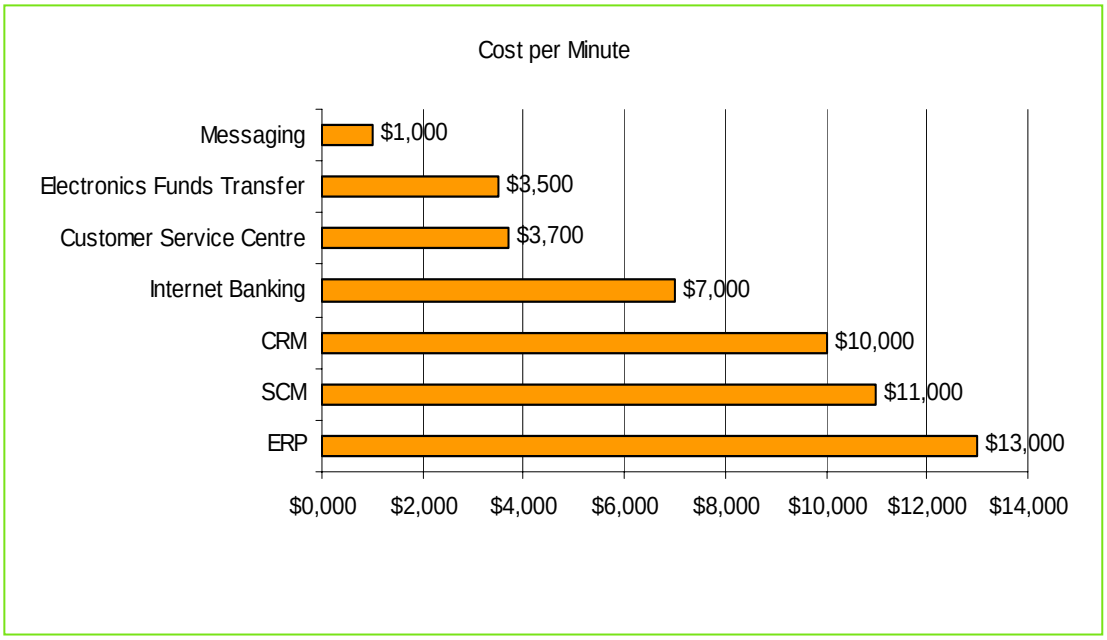


Figure 1 – Average per Minute Cost of Downtime

Source: Standish Group, 2004 - (Study was based upon 250 Fortune 1000 companies with an average of 20,000 users).

CLASSES OF SERVICE

1. Definition
2. CoS benefits and limitations

1. Definition

Service Providers recognize the need to differentiate the network resources offered to different applications; they created Classes of Services or CoS to facilitate this.

The objective of CoS is to adapt the infrastructures behavior to flow types. It is based on the DiffServ model proposed by the IETF (Internet Engineering Task Force), the Internet standardization body, and it uses the following mechanisms:

- Flows are differentiated into a few Classes of Service, generally 3 to 6, called Gold, Silver, Bronze, etc.
- The amount of traffic in higher classes is usually limited, for instance 'no more than 20% of packets can be Gold'. A class overflow may result in packets being discarded or relegated to a lower class.
- The class selection is marked in a field in each packet (DSCP or TOS field). The marking is often done by a router at the customer site or close to it within the network.
- All or some routers in the network will then implement local priority rules based on the value of this field, basically, a packet from a higher class will be routed prior to a packet in a lower class.

It is equivalent to partitioning the initial single-class network into 3 to 6 individual networks with different precedence and performance. A higher class may sometimes borrow bandwidth from others (dependant on the specific telcos implementation).

2. CoS benefits and limitations

2.1 Benefits

CoS addresses the issue of undifferentiated treatment of application flows. By assigning some flows to a higher class, the traffic will spend less time in congested router queues and will have a minimum amount of dedicated bandwidth. As a result, the higher class traffic will have lower transfer delays and lower packet losses.

Real-time traffic like VoIP, along with delay and loss sensitive protocols like SNA over IP, will benefit from having the lowest possible delay within network routers. Critical transactional applications may also benefit from a higher class than that of non-transactional bandwidth hungry applications.

CoS based on the DiffServ model is in theory simple for a service provider to implement inside the network. It requires the implementation of local priority rules in the routers. As such, it doesn't require a lengthy and costly overhaul of the existing network infrastructures.

2.2 Limitations

On the other hand, CoS has some important limitations:

- CoS is focused around infrastructure behavior and does not address end-to-end user needs
- CoS is a best effort based approach to QoS and cannot guarantee the performance of individual critical application flows
- Overall, CoS exacerbates the chronic problem of network over-sizing
- The proper implementation of CoS is complex and time consuming for end-user organizations
- CoS doesn't fit easily into heterogeneous architectures
- CoS leads to rigid implementations
- CoS can be expensive

Focused on infrastructure behavior and does not address end-to-end user needs

A common problem is that there is a difference between a customer expectation's that CoS will improve application performance, and the fact that actual performance depends on both infrastructure performance and the way it is actually used. It is interesting to note that the majority of poor performance helpdesk calls are due to the customer overloading the network rather than an actual network problem.

SLAs currently offered only deal with **statistical measurement of the infrastructure quality and never with real end-to-end traffic performance.** Measurements are on parts of the network only, like POP to POP, and are generated from router statistics or by test traffic from shadow routers.

Best effort basis with no guarantee

With CoS, traffic is classified by group of applications (aggregate behavior), but no differentiation is made by user session, consequently inside a class the standard TCP rules apply and all flows are handled in the same way. As such, if there are several applications in one class and several users using these applications simultaneously, congestion and deterioration of application performance can still be experienced despite CoS. This explain why the CoS approach to QoS is best effort based; it cannot guarantee the performance of individual application flows.

Network over-sizing

As seen previously, CoS is similar to splitting the network into several networks. To minimize the probability of performance deterioration in higher classes, there is a tendency to over-provision them. As this behavior gets replicated over several classes, CoS exacerbates the chronic problem of network over-sizing. In addition, by splitting the network into several networks, CoS automatically reduces the infrastructures sharing capabilities. Overall, CoS will lead to a less efficient use of bandwidth.

Complex implementation

Rules for priority management of classes have to be implemented in all network routers. This includes measures such as traffic limiting for higher classes to protect the quality of delivered service against customer overloads. As bandwidth allocated to each class may vary from customer to customer, and over time for a given customer, entering and maintaining these rules is a heavy management burden for the service providers.

However, the biggest burden is on the customer organization that needs to predict the traffic for each class. This is much more complex to do than for access as a whole, as the customer needs to define and maintain the following per access point:

- the allocation of traffic flows to the various classes;
- the corresponding sizing of the classes

Doesn't fit easily into heterogeneous architectures

The implementation of CoS depends on each telcos unique engineering rules. As such, heterogeneous architectures requiring multiple peering points are a challenge for CoS implementations. Indeed, each peering point will require adaptation and mapping of CoS rules (number of classes, limits and re-coloring rules etc.). In addition sizing at peering interfaces is hard to define. For these reasons, low-cost access networks relying on ADSL accesses may not be CoS compatible. Accesses may rely on the incumbent service provider who cannot or will not implement other Service Providers CoS rules.

Leads to rigid implementation

The complexity of implementation required by CoS introduces a certain level of rigidity. Due to the risk and complexity of any modifications, CoS has very poor flexibility.

Indeed, rolling out a new application or changing the CoS of an existing application will also require changes in many rules; a task that is tedious, long and error prone. As such, there is a natural tendency for client organizations and Service Providers alike to delay or ignore the changes necessary for optimal use of the network. This is unfortunate as networks are "living" bodies with constant changes in terms of user behavior, new application rollout etc.

Can be expensive

Service providers charge CoS at a premium because:

- it offers the customer a new service with new SLAs;
- it requires more powerful network equipment;
- it increases the Service Provider's management burden;
- it usually increases the network size both for access lines and in the core

Customer organizations should expect an increase of between 10% and 20% of their network budget. In addition, any changes to CoS rules are charged to the Customer.

IPANEMA'S BUSINESS OPTIMIZED NETWORKS

The Ipanema System is a network optimization system that in essence provides a real-time and flexible Class of Service per user session.

1. **How does it do that?**
2. **Ipanema System benefits and limitations**

1. How does it do that?

The Ipanema System is made up of a central management software platform, physical and "virtual" devices. Depending on network configuration a combination of physical and virtual devices are deployed at the edges of the network in the relevant branch offices and data centers. The network devices and the central management software form **a distributed Autonomic Networking System** that delivers the visibility, optimization and rightsizing features.

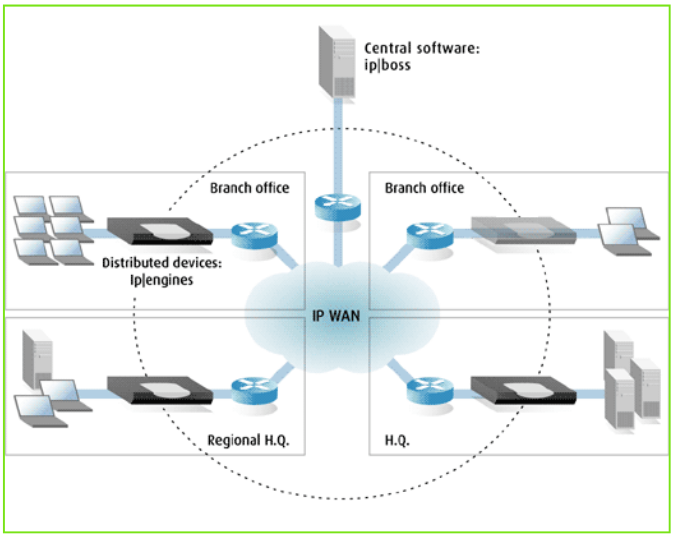


Figure 2. The Ipanema System

The essence of the Ipanema System is straightforward and can be described in three steps:

Firstly, a user ranks applications and user groups (subnets) per business criticality – from high to low- then they determine the performance objectives expected for each application and user group.

Secondly, the System automatically and "autonomically" optimizes network resources to match the performance required. It guarantees application performance for critical applications whatever the

network conditions.

Thirdly, the System provides real-time information about application performance over the network. If non-critical applications do not reach their performance objectives 100% of the time, it provides precisely the amount of bandwidth required to achieve different levels of performance. In other words it allows accurate rightsizing of the network according to application performance objectives or application SLA.

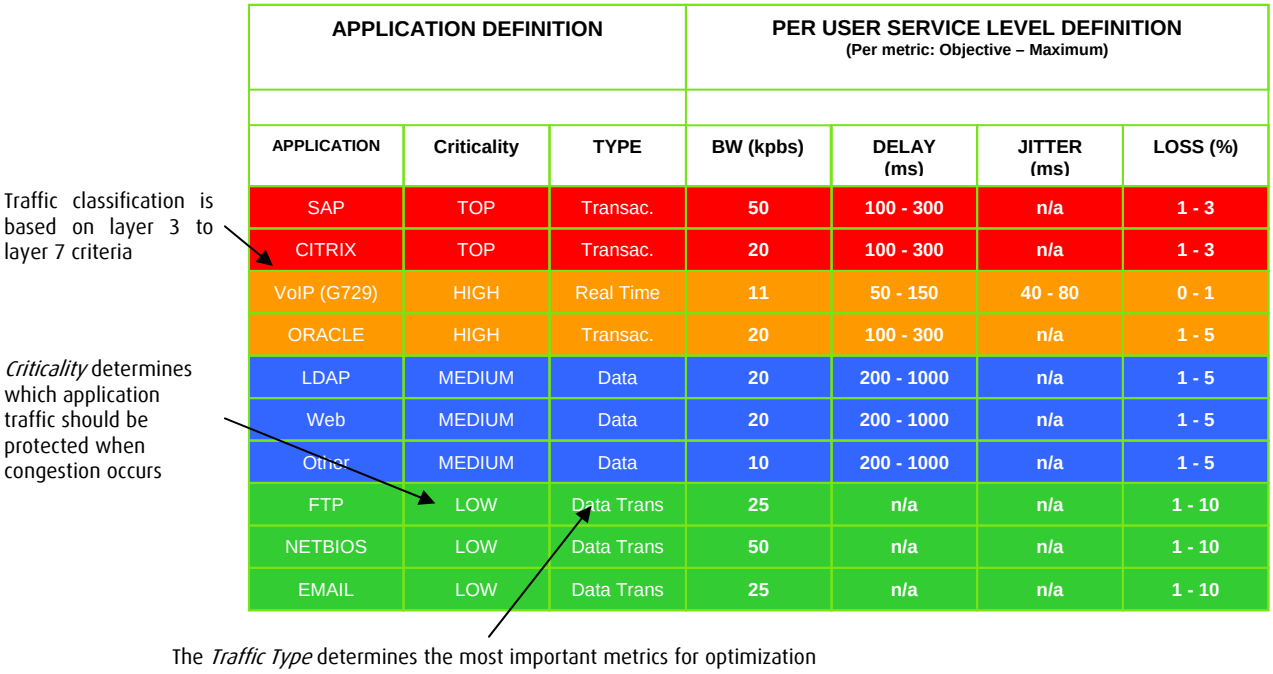
Let's have a more comprehensive look at each step:

Step 1: Defining application performance objectives

Whereas CoS requires modeling of each application, with an allocation of bandwidth required per class and per network access, the Ipanema System only deals with performance objectives at the user level. The client organization will have to determine the business criticality of the application flow and the application performance expected.
Business criticality is voluntarily limited to four levels, and it is straightforward: Top, High, Medium and Low. Expected performance is expressed in terms of minimum required bandwidth, jitter, delay and loss. Both expected performance and business criticality can be expressed at the application level (e.g. SAP critical for all users) and at the user level (e.g. E-mail non-critical for the organization except for traders).

In most cases, 20 to 30 parameters suffice to pilot a network with tens of applications, hundreds of sites and thousands of users.

Figure 3 shows how these parameters are set:


Traffic classification is based on layer 3 to layer 7 criteria
Criticality determines which application traffic should be protected when congestion occurs
The *Traffic Type* determines the most important metrics for optimization

APPLICATION DEFINITION			PER USER SERVICE LEVEL DEFINITION (Per metric: Objective – Maximum)			
APPLICATION	Criticality	TYPE	BW (kpbs)	DELAY (ms)	JITTER (ms)	LOSS (%)
SAP	TOP	Transac.	50	100 - 300	n/a	1 - 3
CITRIX	TOP	Transac.	20	100 - 300	n/a	1 - 3
VoIP (G729)	HIGH	Real Time	11	50 - 150	40 - 80	0 - 1
ORACLE	HIGH	Transac.	20	100 - 300	n/a	1 - 5
LDAP	MEDIUM	Data	20	200 - 1000	n/a	1 - 5
Web	MEDIUM	Data	20	200 - 1000	n/a	1 - 5
Other	MEDIUM	Data	10	200 - 1000	n/a	1 - 5
FTP	LOW	Data Trans	25	n/a	n/a	1 - 10
NETBIOS	LOW	Data Trans	50	n/a	n/a	1 - 10
EMAIL	LOW	Data Trans	25	n/a	n/a	1 - 10

Figure 3. Setting the parameters

Step 2. The System "Autonomically" optimizes network resources

The Ipanema System's optimization is objective driven, global and combines multiple mechanisms :

Objective driven

Ipanema's optimization is not based on static policies based on a snapshot of network conditions. Instead, the Ipanema System constantly adapts to network conditions to meet application performance objectives. These objectives, as mentioned earlier, specify the minimum network requirements for each end-user to get the right quality of experience. In addition, the optimization mechanism uses the notion of business criticality to protect critical applications in case of congestion.

Global

Whereas CoS applies priority rules locally, Ipanema optimizes performances on a network-wide basis. In real-time it manages all types of congestion: locally from the source, at the destination - due to conflicts between traffic coming from several sources - and even network-hosted congestion.

This guarantees that performance objectives are met even in complex networks with a complex traffic matrix. This is a mandatory requirement when several data centers are running simultaneously, or with meshed traffic (VoIP).

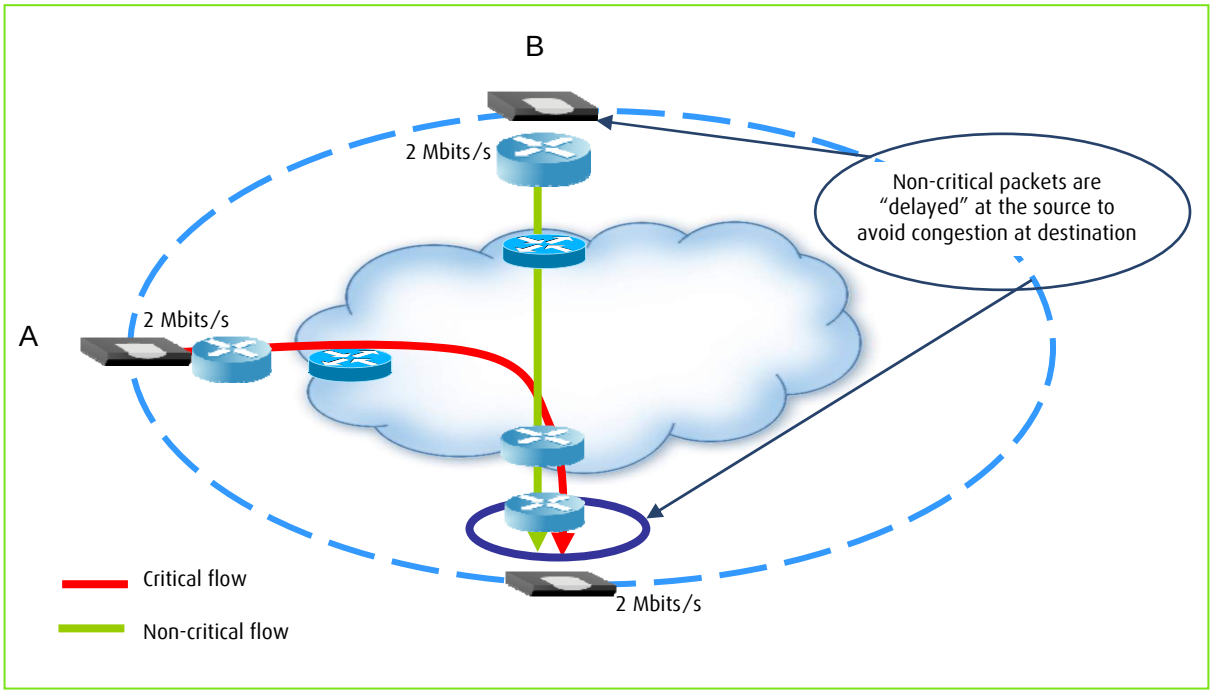


Figure 4. Meshed WAN traffic requires global optimization

Combining multiple mechanisms

The Ipanema System uses three optimization mechanisms that work together; Dynamic Bandwidth Allocation (DBA), Smart Packet Forwarding (SPF) and Advanced Compression. By doing so, it assures that the network always delivers the best possible application performance according to the business objectives. Ipanema's Advanced Compression creates additional bandwidth and increases the link's speed. The Dynamic Bandwidth Allocation (DBA) mechanism insures that the additional resources are used by business critical applications instead of recreational ones. And the Smart Packet Forwarding (SPF) features protect application delays and jitter.

Dynamic Bandwidth Allocation

Whereas CoS allocates static bandwidth per access to each of the classes, Ipanema allocates bandwidth dynamically to each active user session according to:

- the objectives defined above;
- the real-time measurement of available bandwidth between all access points;
- the real-time tracking of the traffic demand.

During congestion, Ipanema ensures that each critical flow gets its bandwidth objectives, while other flows will be allocated a share of the remaining bandwidth according to their relative business criticality.

Smart Packet Forwarding

While the DBA allocates bandwidth according to business criticality and network conditions, the "Smart Packet Forwarding" (SPF) feature decides, in real-time, within allocated bandwidth and dependant on the nature of the flow – i.e. real-time, transactional, data transfer – which packet to "forward first". Thus, the constraints of application delay and jitter sensitivity are protected.

Advanced Compression

Ipanema's Advanced Compression is based on intra-packet and inter-packet, redundancy elimination. By reducing the bandwidth used by the application flows, it creates additional bandwidth and increases a link speed. Configuration of the compression tunnels is fully automated even in meshed and redundant environments. Furthermore this feature is fully compatible with MPLS mechanisms. While compression is very useful and powerful on specific links, it is rarely required across the whole network.

Step 3. Rightsizing according to application performance objectives

Unlike CoS, which leads to over provisioning, with the Ipanema System links can be rightsized. Indeed, the System provides information that directly links bandwidth to application performance. Clients can therefore make informed trade-offs between cost and performance. In the following example, 768Kbps would have been sufficient for a 100% SLA for critical applications; however 1500 kbps would have been necessary if all applications were to achieve their performance objectives 100% of the time.

2. Ipanema System benefits and limitations

2.1 Limitations

The main limitation of the Ipanema System is that it does not intervene in the core network. It takes the core network conditions as a given. Therefore, if performance issues are driven by shortcomings in the core, the Ipanema System will not be able to improve the situation. Having said that, it must be noted that in most cases core networks are designed to handle the aggregate traffic from numerous customers without congestion. Because of their high bandwidth, their contribution to overall traffic delay is generally small. POP-to-POP delays inside a continent will generally be guaranteed to be a few tens of milliseconds, with very few losses.

The main source of congestion, loss and delay comes from the big imbalances between:

1. LAN and access network bandwidth at the source,
2. Core and access network bandwidth at the destination.

2.2 Benefits

Aligns networks resources with business criticality of applications and guarantees their performance from an end user perspective

Unlike CoS, which allocates traffic flows per group of applications, the Ipanema System "shapes and protects" flows per session. As such, it can guarantee the performance of critical applications as requested by the end user. While CoS is a best effort – infrastructure – centric approach to QoS, the Ipanema System is a user centric guarantee.

- **Easy to implement and manage change seamlessly**

Being objective driven Ipanema's performance optimization is extraordinarily simple to manage. The most complex network needs no more than 30 parameters to be managed.

Unlike router-based CoS, there is no need to determine, configure and implement local traffic policies. Thus, rolling-out a new application or adapting to changes in network configuration are all dealt with just a couple of clicks of a mouse.

- **Fully flexible and adaptive to a variety of traffic situations**

Again, the objective driven nature of the Ipanema System and its optimization mechanisms allow it to constantly and instantaneously adapt to changes in traffic situations. There is no need to re-configure new policies if new critical applications have been added or if the traffic behavior from some specific sites has been permanently or temporarily changed. The System will automatically adapt to all these changes. Quite a difference from the rigid rules-based approach of CoS!

- **Can be deployed over heterogeneous networks**

Unlike CoS, the Ipanema System doesn't rely on changes in routers or Service Provider specific engineering rules. By using its own devices located at the edge of the network, the Ipanema System can optimize the use of network resources whatever the underlying network or networks. Multiples heterogeneous networks are no barrier to the System's efficiency.

- **Allows rightsizing of the network**

Whereas the local and static traffic model of router based CoS requires much over sizing to manage uncertain traffic situations, Ipanema's dynamic and global optimization always makes best use of available bandwidth. The need for over-provisioning is completely eliminated as high critical, real-time or transactional flows are individually protected, not only from lower criticalities but from bandwidth hungry flows.

In addition, the rightsizing feature provides detailed information about the optimum network required for different levels of application performance, taking guesswork out of capacity planning.

- **Reduces network budgets**

Unlike CoS, the Ipanema System reduces network related costs. The System has both indirect and direct financial impacts.

It contributes:

- to increasing the overall organization's productivity by eliminating brown-outs,

- to lowering network incident management costs by reducing the number of incidents and the time to repair
- and to decreasing network change management costs.

But, most importantly, the Ipanema System allows the reduction and/or control of telecom costs. It does so by eliminating the need for bandwidth over-provisioning, by reducing bandwidth requirements for applications thanks to optimization factors of up to 400% and finally by affecting end-user behavior.

It is not unusual that due to the Ipanema System, clients can sustain increases of 400% of traffic over several years without requiring any bandwidth increase while guaranteeing application performance. (For more on the financial impact of the Ipanema System refer to the ROI White paper).

SUMMARY

The table below summarizes the main differences between traditional router based CoS and optimization of network resources using the Ipanema System.

Feature	Classes of Service: A network centric approach	Ipanema System: A user centric approach
Management of application performance objectives	Policy driven: applications are grouped in classes	Objective driven: per application at user experience level
Application performance guarantee	No, best effort basis	Yes for business critical applications
Handling of individual user application performance	No, aggregation of all flows from the same class	Yes per user session
SLA	Network centric	User & application centric
Ease of implementation	Low	High
Flexibility	Low	High
Heterogeneous network (ADSL, multi-Telco's, etc.)	Difficult to impossible to implement	Not an issue
Selection of class of service	At each access site, simplified classification	Centrally, network wide, detailed classification
Change Management	Difficult	Seamless
Over provisioning	Increased	Eliminated
Network costs	Increased	Decreased
Impact on core network	Limited	None

IPANEMA AND CoS

Based on the merits and limitations of CoS and Ipanema's Business Optimized Network solution, what should clients do? Choose just one approach or adapt both? Can they co-exist on a network?

The Ipanema System and CoS can co-exist on a network. The Ipanema System is compatible with the CoS and it can even mark the packets for their class.

It is our recommendation that clients take limited CoS; for instance two classes: one for real-time traffic (VoIP, Video) and one for all other types of traffic in conjunction with the Ipanema System. By doing so, they will have the best of both worlds. They will protect themselves from eventual shortcomings of their service providers core networks while benefiting from the dynamic objective driven optimization of the Ipanema System.

Multiplying the number of Classes of Service is on the other hand complex, counter-productive and expensive.

The greater the number of Classes of Service, the higher the complexity of implementation

Dividing bandwidth into several sub-bandwidths reduces the amplitude of bandwidth sharing. Beyond a certain number of classes this phenomenon offsets the benefits of CoS unless all classes are thoroughly over-provisioned.

The higher the number of classes, the higher the overall bandwidth over-provisioning and the resulting network costs.

http://www.ipanematech.com

Worldwide Headquarters:
Ipanema Technologies
28, rue de la Redoute
92260 Fontenay-aux-Roses
FRANCE
Phone: + 33 1 55 52 15 00
Fax: + 33 1 55 52 15 01
E-mail: info.fr@ipanematech.com

United States
Ipanema Technologies Corp.
300 Fifth Avenue,
3rd Floor,
Waltham, MA 02451
Phone: (1) 781 419 6526
Fax: (1) 781 419 6527
E-mail:info.us@ipanematech.com

Germany
Ipanema Technologies GmbH
Gustav-Stresemann-Ring 1
65189 Wiesbaden
GERMANY
Phone: + 49 611 97774 285
Fax: + 49 611 97774 111
E-mail: info.de@ipanematech.com

UK
Ipanema Technologies Ltd
Abbey House
Wellington Way
Weybridge
Surrey, KT13 0TT
UNITED KINGDOM
Phone: + 44 (0)1932 268 380
Fax: + 44 (0)1932 268 381
E-mail: info.uk@ipanematech.com

Copyright © 2007, Ipanema Technologies - All rights reserved

Ipanema and the Ipanema logo are trademarks of Ipanema Technologies.
Trademarks and trade names mentioned herein refer to entities claiming those marks and names or to their products. Ipanema Technologies disclaims any proprietary interest in any trademarks and trade names other than its own.